DataNet Full Proposal: DataNet Federation Consortium

PI: Reagan W. Moore[1], Lead Institution: University of North Carolina at Chapel Hill; Co-PIs: Jon Goodall[9], John Orcutt[3], Arcot Rajasekar[1], William Regli[2]; Senior Personnel: Stanley Ahalt[1], Chitta Baral[5], Kenneth Bollen[1], Ryan Boyles[8], Andrea Chiba[3], Jonathan Crabtree[1], Ken Galluppi[1], Jose-Marie Griffiths[1], Cal Lee[1], Jorge Lobo[11], Julian Lombardi[4], Marilyn Lombardi[1], Richard Marciano[1], Sarah Michalak[1], Thomas Palmeri[7], Art Pasquinelli[10], Sudha Ram[6], Peter Robinson[1], Paul Sheldon[7], Helen Tibbo[1], Michael Wan[3]. [1]University of North Carolina at Chapel Hill (DICE, Libraries, Odum, RENCI, SILS), [2]Drexel University, [3]University of California, San Diego, [4]Duke University, [5]Arizona State University, [6]University of Arizona, [7]Vanderbilt University, [8]North Carolina State University, [9]University of South Carolina, [10]Sun Microsystems, [11]IBM,

Intellectual Merit: The DataNet Federation Consortium (DFC) will implement a policy-driven national data management infrastructure that addresses both the science and engineering data life cycle and the sustainability of data collections and repositories. The motivation for building the DFC comes from the data management requirements from the NSF Science of Learning Centers (EEG / MRI sensor data, video), the NSF Ocean Observatories Initiative (real-time data streams, simulation output, video), the NSF Consortium of Universities for Advancement of Hydrologic Science (point data), the iPlant collaborative (genome databases), the Odum social science institute (statistics), and engineering projects in education and CAD/CAM/CAE archives.

Our approach federates institutional repositories, enabling management of shared collections that are distributed across internet-accessible storage systems. The DFC will build consensus across six principal *communities of practice*: 1) science and engineering projects that drive formation of data collections; 2) facilities & operations center; 3) data cyberinfrastructure technology and research that enables advances in science; 4) policy and standards for data interchange; 5) institutions and sustainability through repurposing of data as reference collections; and 6) outreach and education initiatives that promote student access to collections.

Each community of practice is itself a federation across academic institutions, regional consortia, state institutions, federal institutions, and international collaborations. We view federation as a socialization process that develops consensus on standards for data management policies. By differentiating data life cycle changes as evolution of management policies, it is possible to build generic infrastructure that promotes long-term sustainability. We are already building the data life cycle management infrastructure through integration of the integrated Rule-Oriented Data System (iRODS) with workflow systems, digital libraries, preservation environments, social networking tools, and education tools. We will promote sustainability by enabling multiple communities to use, "own", and store local copies of a data collection, while enforcing global preservation policies. We are collaborating with federal agencies, vendors, and international projects on the development of reference implementations of generic data management infrastructure for use in data sharing, analysis, publication, and preservation.

Broader Impacts: We will promote use of research and engineering collections in education through integration of appropriate analysis, governance, and publication policies, and give workshops and summer school sessions on application of the DFC technology. The ability to involve students in research on "live" data, reinforced with research results from the Science of Learning Centers, can revolutionize interest in science. The ability to federate across existing collections will make it possible to build collections that span institutions, regions, and agencies enabling participation by local projects in national initiatives. The ability to compare archived results with real-time observations will drive new modes of research that dynamically control remote sensors. The integration across multiple communities of practice ensures representation from all stakeholders in the data life cycle. The DFC will implement a preservation environment that facilitates and encourages re-use of data for both research as well as education.